



Using OpenMP from C++ Victor Eijkhout TACC training 2023



Eijkhout: OMP C++

Justification

OpenMP has the opportunity to exploit features of modern C++ that are not present in C. In this course we will explore:

- range-based iteration,
- differences in treatment between vectors and arrays, and various sophisticated reduction schemes.

TACC

Eijkhout: OMP C++

Basic stuff



1. Output streams in parallel

The use of *cout* may give jumbled output: lines can break at each <<.

Use stringstream to form a single stream to output.

```
// hello.cxx
#pragma omp parallel

{
   int t = omp_get_thread_num();
   stringstream proctext;
   proctext << "Hello world from " << t << endl;
   cerr << proctext.str();
}</pre>
```

2. Parallel regions in lambdas

OpenMP parallel regions can be in functions, including lambda expressions.

```
const int s = [] () {
int s;
    # pragma omp parallel
    # pragma omp master
    s = 2 * omp_get_num_threads();
    return s; }();
```

('Immediately Invoked Function Expression')

3. Dynamic scope for class methods

Dynamic scope holds for class methods as for any other function:

Code:

Output:

```
// nested.cxx
class c {
public:
  void f() {
    cout
      <<
    →omp_get_num_threads()
      << '\n':
 };
int main() {
  c my_object;
#pragma omp parallel
  my_object.f();
```

4. Privatizing class members

Class members can only be privatized from (non-static) class methods:

```
class foo {
private:
int x;
public:
void f() {
#pragma omp parallel private x
f g()
}
}
```

So f can not be static, and

```
class foo { public: int x; }
foo x;
#pragma omp parallel private thing.x // NOPE
```

5. Vectors are copied, unlike arrays, 1

```
C arrays: private pointer, but shared array:
Code:
// alloc.c
int *array =
  (int*) malloc(nthreads*sizeof(int));
for (int i=0; i<nthreads; i++)</pre>
  array[i] = 0;
#pragma omp parallel firstprivate(array)
  int t = omp_get_thread_num();
  array += t;
```

Output:

```
1 Array result:
2 0:0, 1:1, 2:2, 3:3,
```

array[0] = t;

// ... print the array

6. Vectors are copied, unlike arrays, 2

C++ vectors: copy constructor also copies data:

Parallel loops

7. Questions

- Do regular OpenMP loops look different in C++?
- 2 Is there a relation between OpenMP parallel loops and iterators?
- OpenMP parallel loops vs parallel execution policies on algorithms.



Eijkhout: OMP C++

8. Range syntax

Parallel loops in C++ can use range-based syntax as of OpenMP-5.0:

```
// vecdata.cxx
   vector<float> values(100):
  #pragma omp parallel for
   for ( auto& elt : values ) {
   elt = 5.f;
   float sum{0.f};
   #pragma omp parallel for reduction(+:sum)
   for ( auto elt : values ) {
     sum += elt;
12
13
```

Tests not reported here show exactly the same speedup as the C code.

9. General idea

OpenMP can parallelize any loop over a C++ construct that has a 'random-access' iterator.



10. C++ ranges header

The C++20 ranges library is also supported:

```
// range.cxx
         pragma omp parallel for reduction(+:count)
         for ( auto e : data )
           count += e;
         pragma omp parallel for reduction(+:count)
         for ( auto e : data
                  | std::ranges::views::drop(1) )
           count += e;
         pragma omp parallel for reduction(+:count)
         for ( auto e : data
10
                  | std::ranges::views::transform
11
                  ([](auto e) { return 2*e; }))
12
           count += e:
13
```

11. C++ ranges speedup

```
==== Run range on 1 threads ====
   sum of vector: 50000005000000 in 6.148
   sum w/ drop 1: 50000004999999 in 6.017
   sum times 2 : 100000010000000 in 6.012
   ==== Run range on 25 threads ====
   sum of vector: 50000005000000 in 0.494
   sum w/ drop 1: 50000004999999 in 0.477
   sum times 2 : 100000010000000 in 0.489
   ==== Run range on 51 threads ====
   sum of vector: 50000005000000 in 0.257
10
   sum w/ drop 1: 50000004999999 in 0.248
11
   sum times 2 : 100000010000000 in 0.245
12
   ==== Run range on 76 threads ====
13
   sum of vector: 50000005000000 in 0.182
14
15
   sum w/ drop 1: 50000004999999 in 0.184
   sum times 2 : 100000010000000 in 0.185
16
   ==== Run range on 102 threads ====
   sum of vector: 50000005000000 in 0.143
18
   sum w/ drop 1: 50000004999999 in 0.139
19
   sum times 2 : 100000010000000 in 0.134
20
   ==== Run range on 128 threads ====
22;ikhosumpofcyector: 50000005000000 in 0.122
```

TACC

12. Custom iterators, 0

Recall that

Short hand:

```
1 vector<float> v;
2 for ( auto e : v )
3 ... e ...
```

for:

```
for ( vector<float>::iterator
e=v.begin();
e!=v.end(); e++ )
... *e ...
```

If we want

```
1 for ( auto e : my_object )
2    ... e ...
```

we need a sub-class for the iterator with methods such as begin, end, * and ++.

13. Custom iterators, 1

OpenMP can parallelize any range-based loop with a random-access iterator.

Class:

```
1 // iterator.cxx
2 template<typename T>
3 class NewVector {
4 protected:
5    T *storage;
6    int s;
7 public:
8    // iterator stuff
9    class iter;
10    iter begin();
11    iter end();
12 };
```

Main:

```
NewVector<float> v(s);
pragma omp parallel for
for ( auto e : v )
cout << e << " ";</pre>
```

14. Custom iterators, 2

Required iterator methods:

```
NewVector<T>::iter& operator++();
T& operator*();
bool operator==( const NewVector::iter &other ) const;
bool operator!=( const NewVector::iter &other ) const;
// needed to OpenMP
int operator-( const NewVector::iter& other ) const;
NewVector<T>::iter& operator+=( int add );
```

This is a little short of a full random-access iterator; the difference depends on the OpenMP implementation.

TACC

15. Custom iterators, exercise

Write the missing iterator methods. Here's something to get you started.

```
template<typename T>
   class NewVector<T>::iter {
   private: T *searcher;
   template<typename T>
   NewVector<T>::iter::iter( T* searcher )
     : searcher(searcher) {}:
   template<typename T>
   NewVector<T>::iter NewVector<T>::begin() {
     return NewVector<T>::iter(storage); };
10
   template<typename T>
11
   NewVector<T>::iter NewVector<T>::end()
12
     return NewVector<T>::iter(storage+NewVector<T>::s); };
13
```

16. Custom iterators, solution

```
template<typename T>
   bool NewVector<T>::iter::operator==( const NewVector<T>::iter &other
        \hookrightarrow) const {
     return searcher==other.searcher: }:
   template<typename T>
   bool NewVector<T>::iter::operator!=( const NewVector<T>::iter &other
        \hookrightarrow) const {
     return searcher!=other.searcher; };
   template<typename T>
   NewVector<T>::iter& NewVector<T>::iter::operator++() {
     searcher++; return *this; };
   template<typename T>
10
   NewVector<T>::iter& NewVector<T>::iter::operator+=( int add ) {
11
     searcher += add; return *this; };
12
```

17. Custom iterators, solution

TACC

18. OpenMP vs standard parallelism

Application: prime number marking (load unbalanced)

Standard parallelism uses Thread Building Blocks (TBB) as backend

19. Timing

```
==== Run primepolicy on 1 threads ====
  OMP: found 0 primes; Time: 390 msec (threads= 1)
   TBB: found 0 primes; Time: 392 msec
   ==== Run primepolicy on 25 threads ====
  OMP: found 0 primes; Time: 17 msec (threads=25)
  TBB: found 0 primes; Time: 19 msec
   ==== Run primepolicy on 51 threads ====
  OMP: found 0 primes; Time: 9 msec (threads=51)
  TBB: found 0 primes; Time: 13 msec
10
   ==== Run primepolicy on 76 threads ====
   OMP: found 0 primes; Time: 6 msec (threads=76)
11
  TBB: found 0 primes; Time: 15 msec
12
   ==== Run primepolicy on 102 threads ====
13
   OMP: found 0 primes; Time: 5 msec (threads=102)
14
  TBB: found 0 primes; Time: 71 msec
15
   ==== Run primepolicy on 128 threads ====
16
  OMP: found 0 primes; Time: 4 msec (threads=128)
17
  TBB: found 0 primes; Time: 55 msec
18
```

20. Reductions vs standard parallelism

Application: prime number counting (load unbalanced) missing snippet reduceprimeomp missing snippet reduceprimecpp



21. Timing

```
==== Run reducepolicy on 1 threads ====
                                    390 msec (threads= 1)
   OMP: found 9592 primes; Time:
   TBB: found 9592 primes; Time:
                                    392 msec
   ==== Run reducepolicy on 25 threads ====
   OMP: found 9592 primes; Time:
                                17 msec (threads=25)
   TBB: found 9592 primes; Time: 20 msec
   ==== Run reducepolicy on 51 threads ====
   OMP: found 9592 primes; Time:
                                   8 msec (threads=51)
   TBB: found 9592 primes; Time:
                                   13 msec
10
   ==== Run reducepolicy on 76 threads ====
11
   OMP: found 9592 primes; Time:
                                   6 msec (threads=76)
   TBB: found 9592 primes; Time:
12
                                   23 msec
   ==== Run reducepolicy on 102 threads ====
13
                                5 msec (threads=102)
   OMP: found 9592 primes; Time:
14
   TBB: found 9592 primes; Time:
                                    105 msec
15
   ==== Run reducepolicy on 128 threads ====
16
   OMP: found 9592 primes; Time:
                                   4 msec (threads=128)
17
   TBB: found 9592 primes; Time:
                                   54 msec
18
```

Reductions



22. Questions

- Are simple reductions the same as in C?
- ② Can you reduce std::vector like an array?
- Precisely what can you reduce?
- 4 Any interesting examples?
- **6** Compare reductions to native C++ mechanisms.

TACC

23. Scalar reductions

Same as in C, you can now use range syntax for the loop.

24. Reductions on vectors

Use the *data* method to extract the array on which to reduce. Also, the reduction clause wants a variable, not an expression, for the array, so you need an extra bare pointer:

25. Reduction on class objects

Reduction can be applied to any class for which the reduction operator is defined as *operator*+ or whichever operator the case may be.

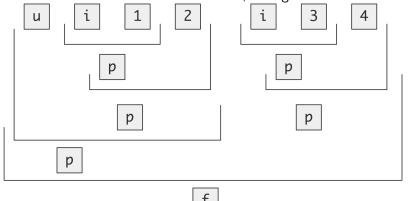
```
// reductcomplex.cxx
                                      vector< Thing >
   class Thing {
                                             \hookrightarrowthings(500, Thing(1.f));
                                        Thing result(0.f);
   private:
     float x;
                                        #pragma omp parallel for
                                             public:
     Thing() : Thing( 0.f ) {};
                                     4 for ( const auto& t : things )
     Thing(float x): x(x) {};
                                       result = result + t;
                                     5
     Thing operator+( const Thing&
       \hookrightarrow other ) {
       return Thing( x + other.x );
    };
10
   };
```

A default constructor is required for the internally used init value; see figure 31.



26. Reduction illustrated

Reduction of four items on two threads, taking into account initial values.



27. User-defined reductions, syntax

```
#pragma omp declare reduction
( identifier : typelist : combiner )
[initializer(initializer-expression)]
```

28. Reduction over iterators

Support for C++ iterators

29. Lambda expressions in declared reductions

You can use lambda expressions in the explicit expression:

```
// reductexpr.cxx
#pragma omp declare reduction\
(minabs : int : \
omp_out = \
[] (int x,int y) -> int { \
return abs(x) > abs(y) ? abs(y) : abs(x); } \
(omp_in,omp_out) ) \
initializer (omp_priv=limit::max())
```

You can not assign the lambda expression to a variable and use that, because omp_in/out are the only variables allowed in the explicit expression.



30. Example category: histograms

Count which elements fall into what bin:

```
for ( auto e : some_range )
    histogram[ value(e)]++;
```

Collisions are possible, but unlikely, so critical section is very inefficient



31. Histogram: intended main program

Declare a reduction on a history gram object:

Q: why does the inc not have to be atomic?

32. Histogram: reduction operator

Give the class a += operator to do the combining:

```
// mapreduce.cxx
   template<typename key>
   class bincounter : public map<key,int> {
   public:
   // merge this with other map
     void operator+=( const bincounter<key>& other ) {
       for ( auto [k,v] : other )
          if ( map<key,int>::contains(k) ) // c++20
            this->at(k) += v:
          else
10
            this->insert(\{k,v\});
11
12
     };
13
   // insert one char in this map
     void inc(char k) {
14
        if ( map<key,int>::contains(k) )
15
          this->at(k) += 1;
16
       else
17
         this->insert(\{k,1\});
18
     };
```

Eijkhout: OMP C++

33. Histogram in native C++

Use atomics because there is no reduction mechanism:

```
// mapreduceatomic.cxx
   class CharCounter : public array<atomic<int>,26> {
   public:
     CharCounter() {
       for ( int ic=0; ic<26; ic++ )
         (*this)[ic] = 0;
   };
   // insert one char in this map
   void inc(char k) {
       if (k==' ') return;
10
       int ik = k-'a';
11
       (*this)[ik]++;
12
    };
13
14 };
```

34. Histogram in native C++, comparison

OpenMP reduction on array<int,26>:

```
1 Using atomics on 1 threads: time= 20.19 msec
2 OpenMP reduction on 1 threads: time= 1.966 msec
3 Using atomics on 5 threads: time= 315.855 msec
4 OpenMP reduction on 5 threads: time= 0.52 msec
5 Using atomics on 10 threads: time= 91.968 msec
6 OpenMP reduction on 10 threads: time= 0.364 msec
7 Using atomics on 30 threads: time= 249.171 msec
8 OpenMP reduction on 30 threads: time= 0.556 msec
9 Using atomics on 50 threads: time= 164.177 msec
10 OpenMP reduction on 50 threads: time= 0.904 msec
```

TACC

39

Eijkhout: OMP C++

35. Example category: list filtering

The sequential code is as follows:

```
vector<int> data(100);
// fil the data
vector<int> filtered;
for ( auto e : data ) {
   if ( f(e) )
   filtered.push_back(e);
}
```

36. List filtering, solution 1

Let each thread have a local array, and then to concatenate these:

```
#pragma omp parallel
{
    vector<int> local;

# pragma omp for
    for ( auto e : data )
        if ( f(e) ) local.push_back(e);

filtered += local;
}
```

where we have used an append operation on vectors:

```
// filterreduct.cxx
template<typename T>
vector<T>& operator+=( vector<T>& me, const vector<T>& other ) {
me.insert( me.end(),other.begin(),other.end() );
return me;
};
```

37. List filtering, not quite solution 2

We could use the plus-is operation to declare a reduction:

Problem: OpenMP reductions can not be declared non-commutative, so the contributions from the threads may not appear in order.

```
Code:
```

```
Output:
```

```
#pragma omp parallel \
    reduction(+ : filtered)

{
    vector<int> local;
    # pragma omp for
    for ( auto e : data )
        if ( f(e) )
        local.push_back(e);
    filtered += local;
}
```

```
→5 10 15 20 25

→30 35 40 45 50

→55 60 65 70 75
```

Mod 5: 80 85 90 95 100

38. List filtering, task-based solution

With a task it becomes possible to have a spin-wait loop:

Code:

Output:

39. Templated reductions

You can reduce with a templated function if you put both the declaration and the reduction in the same templated function:

```
template<typename T>
  T generic_reduction( vector<T> tdata ) {
  #pragma omp declare reduction
  (rwzt:T:omp_out=reduce_without_zero<T>(omp_out,omp_in))
  initializer(omp_priv=-1.f)

  T tmin = -1;
  #pragma omp parallel for reduction(rwzt:tmin)
  for (int id=0; id<tdata.size(); id++)
      tmin = reduce_without_zero<T>(tmin,tdata[id]);
  return tmin;
};
```

which is then called with specific data:

```
1 auto tmin = generic_reduction<float>(fdata);
```



More topics



40. Threadprivate random number generators

The new $C++\ random$ header has a threadsafe generator, by virtue of the statement in the standard that no STL object can rely on global state.

The usual idiom can not be made threadsafe because of the initialization:

```
static random_device rd;
static mt19937 rng(rd);
```

However, the following works:

```
// privaterandom.cxx
static random_device rd;
static mt19937 rng;
#pragma omp threadprivate(rd)
#pragma omp threadprivate(rng)

int main() {

#pragma omp parallel
rng = mt19937(rd());
```

Eijkhout: OMP C++---

You can then use the generator safely and indep

41. Uninitialized containers

Multi-socket systems:

parallel initialization instantiates pages on sockets:

'first touch'

double *x = (double*)malloc(N*sizeof(double));

#pragma omp parallel for
for (int i=0; i<N; i++)
x[i] = f(i);

This does not work with

```
std::vector<double> x(N);
pragma omp parallel for
for (int i=0; i<N; i++)
x[i] = f(i);</pre>
```

because of value initialization in the vector container.

42. Uninitialized containers, 2

Trick to create a vector of uninitialized data:

```
// heatalloc.cxx
  template<typename T>
  struct uninitialized {
  uninitialized() {}:
  T val;
  constexpr operator T() const {return val;};
    T operator=( const T&& v ) { val = v; return val; };
  };
so that we can create vectors that behave normally:
   vector < uninitialized < double >> x(N), y(N);
2
  #pragma omp parallel for
  for (int i=0; i<N; i++)</pre>
  y[i] = x[i] = 0.;
6 x[0] = 0; x[N-1] = 1.;
```

43. Atomic updates

Pragma atomic only works for simple cases. Can you atomically do more complicated updates?

- Make an object that has data plus a lock;
- Disable copy and copy-assignment operators;
- Destructor does omp_destroy_lock;
- Overload arithmetic operator.



Eijkhout: OMP C++

44. Atomic updates: class with OMP lock

```
// lockobject.cxx
   class atomic int {
   private:
     omp_lock_t the_lock;
     int _value{0};
   public:
     atomic int() {
       omp_init_lock(&the_lock);
     };
     atomic_int( const atomic_int& )
10
          = delete:
11
     atomic_int& operator=( const atomic_int& )
12
          = delete:
13
     ~atomic_int() {
14
15
       omp_destroy_lock(&the_lock);
     };
16
```

45. Atomic updates: atomic ops

```
int operator +=( int i ) {
// atomic increment
omp_set_lock(&the_lock);
_value += i; int rv = _value;
omp_unset_lock(&the_lock);
return rv;
};
```

46. Atomic updates: usage

```
1 atomic_int my_object;
2 vector<std::thread> threads;
3 for (int ithread=0; ithread<NTHREADS; ithread++) {
4    threads.push_back
5    ( std::thread(
6        [=,&my_object] () {
7         for (int iop=0; iop<nops; iop++)
8             my_object += 1; } ) );
9 }
10 for ( auto &t : threads )
11    t.join();</pre>
```

47. Atomic updates, comparison to native

Timing comparison on simplest case:

Object with built-in lock:

Native C++ atomics:

```
std::atomic<int> my_object{0};

#pragma omp parallel for

for ( size_t update=0;

update<NTHREADS*nops;

update++) {

my_object += 1;

}

result = my_object;</pre>
```

Native solution is 10x faster.

48. False sharing prevention

```
#include <new>
       #ifdef __cpp_lib_hardware_interference_size
       const int spread = std::hardware_destructive_interference_size
                / sizeof(datatype);
       #else
       const int spread = 8;
       #endif
       vector<datatype> k(nthreads*spread);
10
       #pragma omp parallel for schedule( static, 1 )
11
       for ( datatype i = 0; i < N; i++ ) {
12
         k[(i\%nthreads) * spread] += 2;
13
14
```

49. Beware vector-of-bool!

Does not compile:

```
// boolrange.cxx
vector<bool> bits(1000000);
for ( auto& b : bits )
b = true;
```

More subtle:

```
Code:
// booliter.cxx
vector<bool> bits(3000000);
#pragma omp parallel for schedule(static,4)
for ( int i=0; i<bits.size(); i++ )
  bits[i] = ( i%3==0 );</pre>
```

Output:

- 4 #threads=4; should be →million: 999659

Different bits[i] are falsely shared.

50. CMake

```
cmake_minimum_required( VERSION 3.20 )
   project( ${PROJECT_NAME} VERSION 1.0 )
   find_package(OpenMP)
   if(OpenMP_CXX_FOUND)
   else()
           message( FATAL_ERROR "Could not find OpenMP" )
   endif()
   add_executable( ${PROJECT_NAME} ${PROJECT_NAME}.cxx )
10
   target_compile_features( ${PROJECT_NAME} PRIVATE cxx_std_17 )
11
   target_link_libraries( ${PROJECT_NAME} PUBLIC OpenMP::OpenMP_CXX )
12
13
   install( TARGETS ${PROJECT NAME} DESTINATION . )
14
```